Data Visualization Coursework (SET09120)

Venetsia Krasteva

School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK
40313507@live.napier.ac.uk

Abstract. In this coursework, a dataset is being visualized so patterns and outliers can be found.
The ordinal discrete data group and wind direction were ordered for all graphs so it is easier to see what the pattern is.
There are two classes of outliers found – Height and Distance. There are five patterns found explaining what they mean for the data.
The first relationship is between the Angle and Distance which shows how the distance changes with the angle. The second relationship found is between Score 1 and the experience level showing how the most experienced group has scored. The third relationship shows what is the different offset for the experienced group between the genders. The fourth relationship found is showing depending on the age what score 2 have they gotten depending on the Wind Direction. The fifth relationship shows how high was the throw between the experience level depending on the Gender and Wind Direction.
The background color is always minimal in all graph but you can still see the lines so it is easier to follow the data.
Finally, an evolution on a given chart was made showing the bad aspects of it and what could be done differently so one might understand the chart correctly with no confusion. Visualization has also been provided to show a different more correct way to visualize the data.

1 Outliers

1.1 First Outlier class

A histogram was used to illustrate height as we only have one dimension of continuous data where there are no gaps and it is easier to see big outliers. In the histogram (See **Fig. 1)** the continuous data height is being plotted and it clearly shows some data that differs significantly from other observations which are after a hundred and fifty. The pre-attentive feature enclosure was used to highlight the outlier.

1.2 Second Outlier class

A scatterplot was used to show distance has outliers as the outliers are not that big for a boxplot or a histogram. In the scatterplot (See **Fig.2**) the outliers can easily be identified because of the data points not being in the pattern showing in the graph. Dis-

tance is the outlier because plotted with anything else the dots always appear on the side. The pre-attentive feature enclosure was used to highlight some of the outliers.

2 Relationships

2.1 First Relationship

The scatterplot (See **Fig.3**) illustrates the strong curved quadratic negative relationship between two continuous variables (distance and angle) and how the distance of a throw changes depending on the angle. There are some outliers in the graph because of the distance but they were not removed because they do not decrease the data density.
A line of best fit is shown, and it is clear by that that there is a strong correlation between the data as it directly goes over the points.
The distance is at its max value when the angle is at its median value. Because it is a curved pattern the rate of increase or decrease changes when one variable changes.

2.2 Second Relationship

The box plot (See **Fig. 4**) shows the relationship for two-dimensional data with one discrete (Score 1) and one continuous (the Experience level). You can see that the Very Experienced (V) have a higher score in both genders than any other experience group and it does not depend on Wind Direction.
The pre-attentive feature color was used to show the different experience levels as well as highlight the Very Experienced by making it a brighter color than the other, so it stands out. As a retinal variable for the value the colors I have used the Viridis scale which is based on luminance to minimize issues with the colors.

2.3 Third Relationship

A box plot was used in juxtaposition so it can show the different types of experience levels and divide them by gender because one dimension is discrete and one continuous.
This box plot (See **Fig. 5**) shows the relationship between the experience level and offset where the females are throwing closer to the centre when they are more experienced where the males do not change in offset no matter what their experience level is.
The law of similarity was used to show the two different types of gender where they are each in a specific color.

2.4 Forth Relationship

This scatterplot (See **Fig. 6**) shows two strong linear relationships between the age and score 2 depending on the Wind Direction but only on East and West. The relationship depending on the East is a positive one where the score is getting higher if

the person is older. The relationship on West is negative as Score 2 gets lower when the person is getting older.
The retinal variable shape is being used as a secondary indicator to show the different types of wind direction.
As a pre-attentive feature, the color of different shades of grey was used to represent the wind directions that do not show us any valuable information whereas the different shades of red were used to highlight both of the relationships so they outstand and are the most visible because of they are not confounding features.

2.5 Fifth Relationship

A box plot was used in a juxtaposition so it can visualize the height depending on the experience level for both genders in the four wind directions.
This boxplot (See **Fig. 7**) illustrates how the males depending on their experience level have a higher height on the north wind direction and a lower height on the south wind direction. The females do not change the height of their throw no matter what their experience level is or the wind direction. Here the outliers ware removed for data density purposes.
The law of common region was applied by the pre-attentive feature enclosure to divide the two genders and make it obvious they are different. In addition, color as a pre-attentive feature is being used to illustrate the different types of wind directions in the Viridis color scale. The law of proximity is used to show the different types of experience level.

3 Visualization Evaluation

The visualization provided for evaluation has been evaluated into different components and provides pros and cons.
There sketched graph by me (See **Fig. 8**) where all the cons have been considered and it was illustrated differently. Firstly the bar chart has the x-axis as countries but they are ordered in alphabetical order and there is a label for it. To the y-axis, a label was added, to sum up, that it is representing the continuous data height. The title of the graph has been put on top of it. Representation of the columns has changed both with color, that represents each country on its own using the Viridis color scale and represented in alphabetical order. A legend has also been provided for the graph to show the different colors for the countries that have been included in the x-axis.
Evaluation:

1. Data Component - The data visualization is using nominal discrete data for the x-axis which does not mean they should be in order, but it may be a good idea to put them alphabetically and continuous data is used for the y-axis.
2. Geometric Component: - since the visualization is using one dimension discrete and one continuous which here it is used as a bar chart so we can see the differences in the average female height in the different countries. The lines from the graph have also been removed which causes difficulty in some cases as the columns represented by the female shape are confusing.

3. Labels Component:
    a. Pros:
        i. There is a title on the visualization which is included within the chart itself and it does not compromise the visualization.
    b. Cons:
        i. Additional labels could be added for the x-axis and the y-axis with the description Countries for the x-axis and Average Height for the y-axis.
        ii. There is no legend as to what the colors in it mean which causes difficulty to understand what each of the identical colors means if they were placed on purpose.
4. Retinal Variables:
    a. Shape: the shape that it is chosen to show the average height of females in a specific country is not needed as the title shows what the visualization is illustrating. Some of the female shapes are covering Australia almost completely which may lead you to think that there may be something more. It could also be said that shape is used as a second indicator although there is only one shape for each of the bars which are misleading, so it is not needed.
    b. Color:
        i. Pros: The colors are in sequential order which means it could be safe for a color-blind person to see the difference.
        ii. Cons: The color in the graph is misleading as one color is showing multiple values which breaks the Law of Similarity as one may get confused to as why is Latvia and South Africa the same color or Austria and Peru as they might think they have a connection.

4 Images of Visualizations

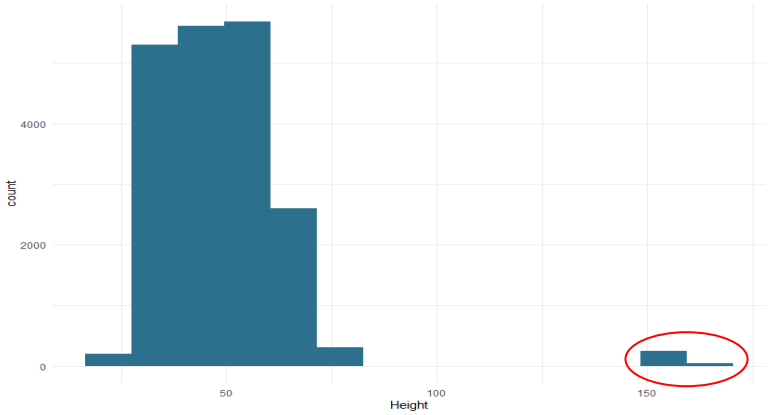

Fig. 2. – Histogram of Height (Outlier)

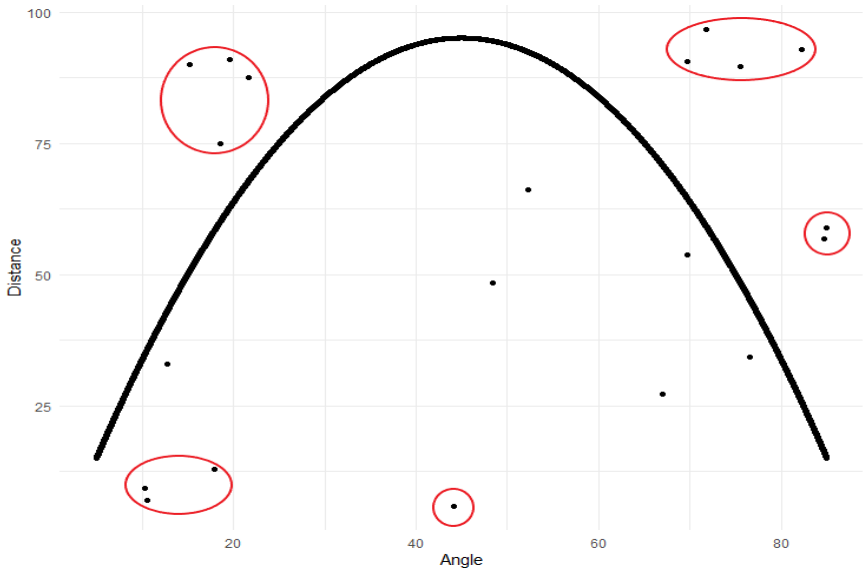

Fig. 2. – Scatterplot showing the outliers in Distance

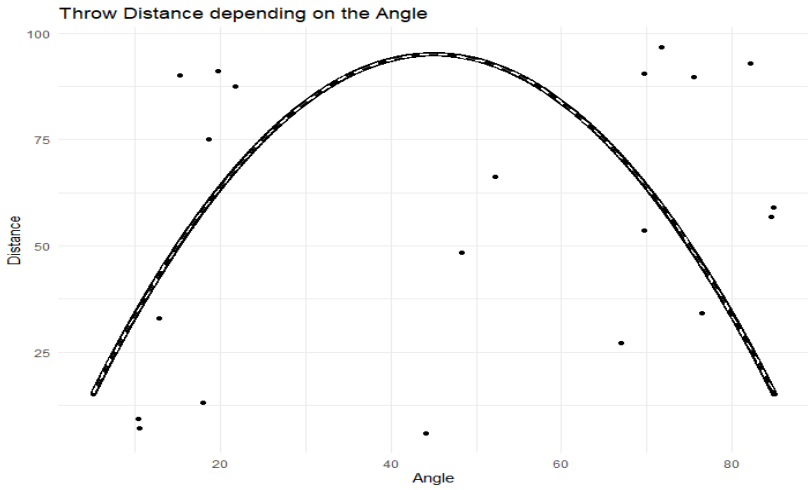

Fig. 3. – Scatterplot of Distance and Angle

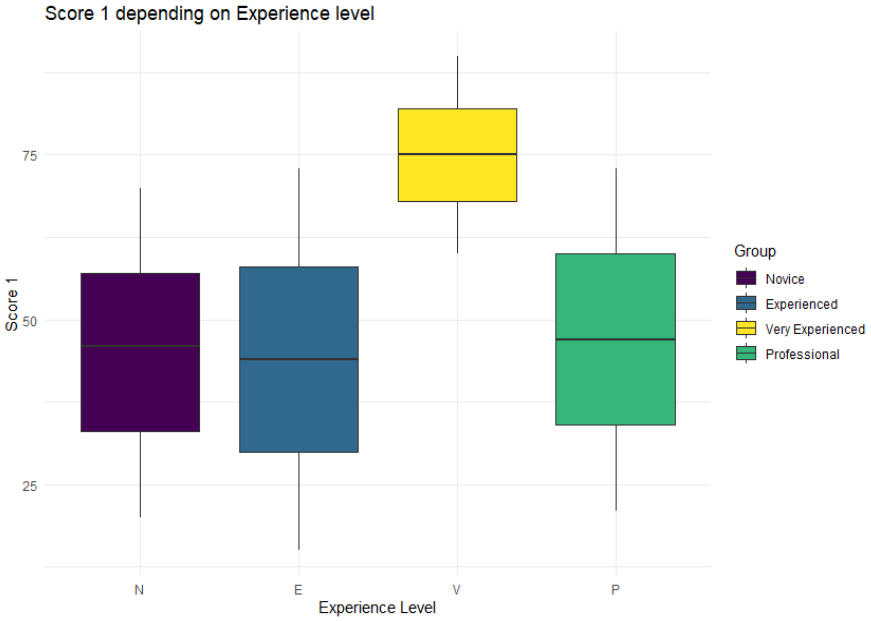

Fig. 4. – Box plot of Score 1 depending on Experience Level

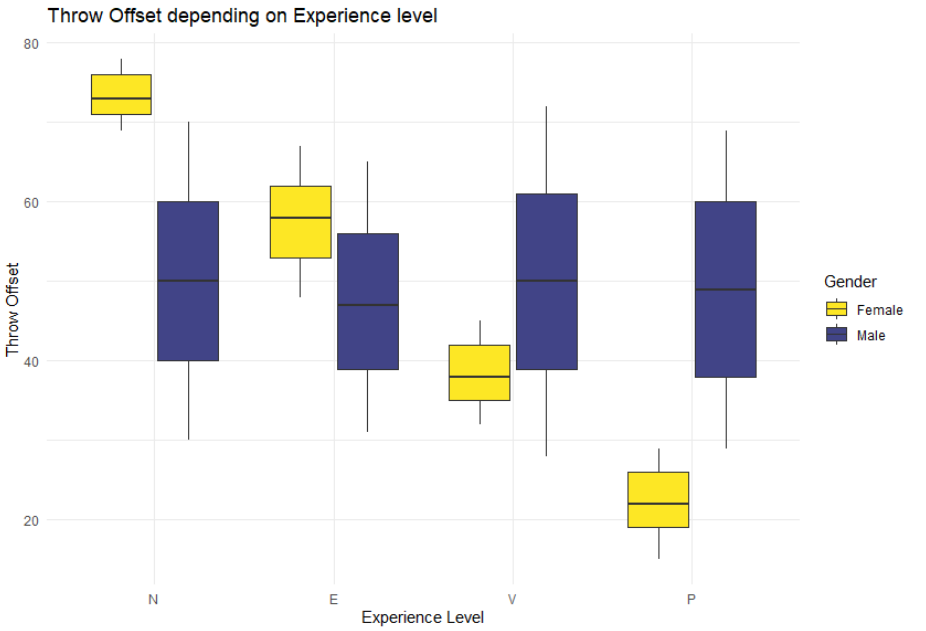

Fig. 5. – Bxplot Offset depending on Experience Level

Fig. 6. - Scatterplot of Score 2 depending on Age categorized by the wind direction

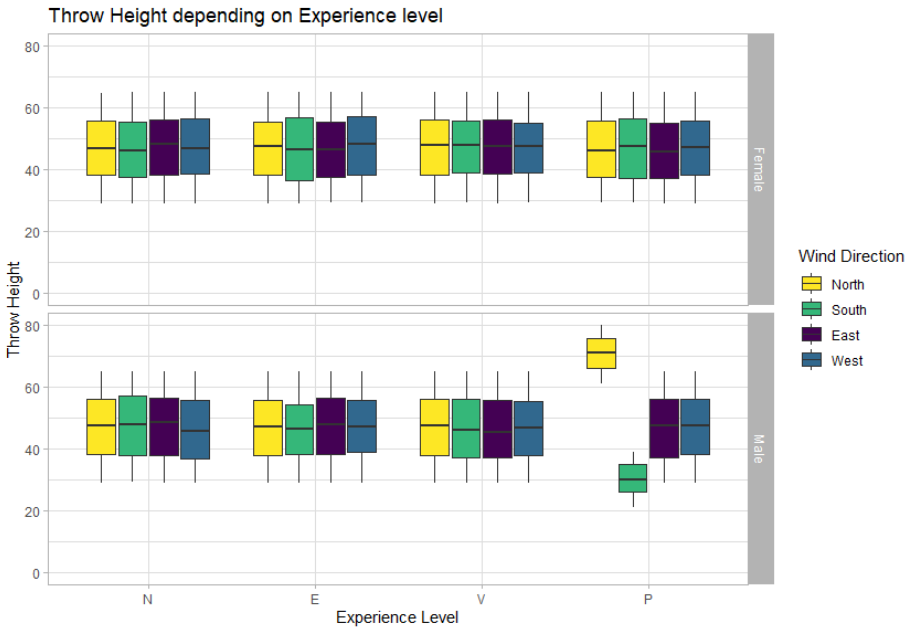

Fig. 7. – Boxplot of Height depending on Experience Level

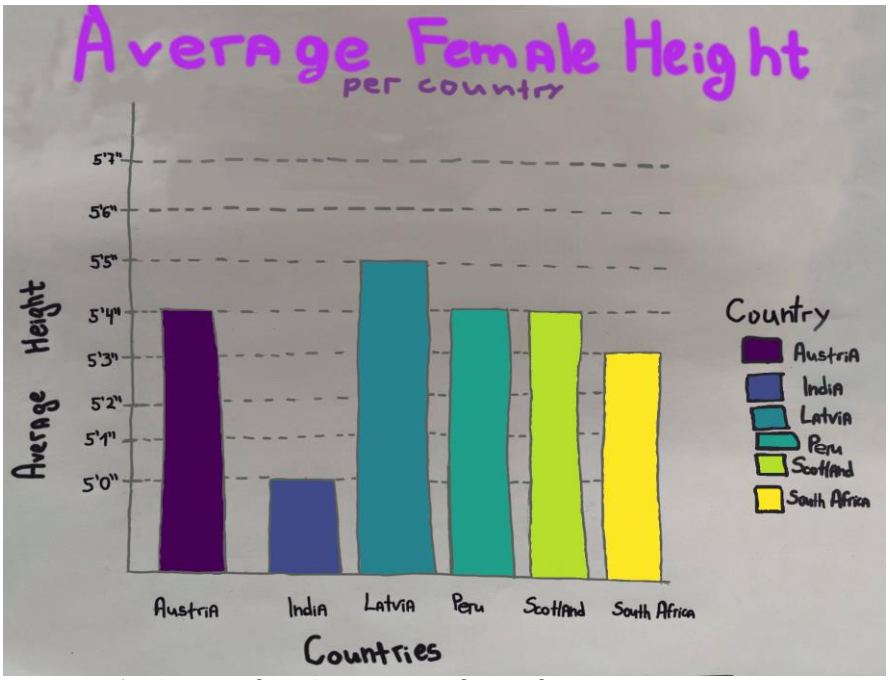

Fig. 8. – Bar chart Average Female Height per country